AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients

Year: 2020 | Citations: 593 | Authors: Juntang Zhuang, Tommy M. Tang, Yifan Ding, S. Tatikonda, N. Dvornek

---

Abstract

Most popular optimizers for deep learning can be broadly categorized as adaptive methods (e.g. Adam) and accelerated schemes (e.g. stochastic gradient descent (SGD) with momentum). For many models such as convolutional neural networks (CNNs), adaptive methods typically converge faster but generalize worse compared to SGD; for complex settings such as generative adversarial networks (GANs), adaptive methods are typically the default because of their stability.We propose AdaBelief to simultaneously achieve three goals: fast convergence as in adaptive methods, good generalization as in SGD, and training stability. The intuition for AdaBelief is to adapt the stepsize according to the "belief" in the current gradient direction. Viewing the exponential moving average (EMA) of the noisy gradient as the prediction of the gradient at the next time step, if the observed gradient greatly deviates from the prediction, we distrust the current observation and take a small step; if the observed gradient is close to the prediction, we trust it and take a large step. We validate AdaBelief in extensive experiments, showing that it outperforms other methods with fast convergence and high accuracy on image classification and language modeling. Specifically, on ImageNet, AdaBelief achieves comparable accuracy to SGD. Furthermore, in the training of a GAN on Cifar10, AdaBelief demonstrates high stability and improves the quality of generated samples compared to a well-tuned Adam optimizer. Code is available at this https URL